CLAIMS

What is claimed is:

1. A method for normalizing text in a document, said method comprising the steps of:

a) generating a list of reference words and phrases and a list of non-reference words and phrases from a selected group of documents;

b) comparing said list of reference words and phrases with a joined list containing said reference words and phrases and said non-reference words and phrases, using an edit-distance algorithm to create an approximate duplicates list;

c) filtering said approximate duplicates list to create a thesaurus of standard words and phrases and their variations; and

d) editing said selected group of documents with an editor operable to use said thesaurus to replace a word or phrase on said approximate duplicates list with said standard words and phrases.

2. The method of Claim 1, wherein words and phrases from said selected group of documents that are on a stop word list are discarded.

3. The method of Claim 2, wherein words and phrases not discarded comprise said lists of reference and non-reference words and phrases.

4. The method of Claim 1, wherein said step of generating further comprises:

a1) counting the frequency of occurrence of a plurality of words and phrases from said selected group of documents;

a2) placing words and phrases with special characters embedded within them on said reference word list;

a3) processing words and phrases from said selected group of documents not already on said reference word list with a spell-checker program, wherein words and phrases that are recognized as correctly spelled are placed on said reference word list and all unrecognized words and phrases are placed on said non-reference word list;

a4) setting a frequency of occurrence threshold for said reference word list, wherein words and phrases which have a frequency of occurrence below said threshold are discarded as irrelevant; and

a5) setting a word frequency threshold for said non-reference word list, wherein words and phrases which have a frequency of occurrence above said threshold remain on said non-reference word list.

5. The method of Claim 4, wherein said reference word list can be merged with an existing domain specific dictionary.

6. The method of Claim 1, wherein said step of comparing comprises:

b1) setting parameters for said edit distance algorithm;

b2) combining said reference word list with said non-reference word list to create a joined list;

b3) comparing words and phrases on said joined list with words and phrases on said reference word list using said edit distance algorithm; and

b4) pairing words and phrases from said non-reference word list with words and phrases from said reference word list, wherein pairs of said words and phrases which are within said parameters of said edit distance algorithm are placed on said approximate duplicates list.

7. The method of Claim 6, wherein said step of comparing further comprises:

setting a parameter, based upon frequency of occurrence, for words and phrases not on said approximate duplicates list;

placing words and phrases not on said approximate duplicates list which are within said parameter on said reference word list; and

discarding words and phrases not on said approximate duplicates list which are outside said parameter.

8. The method of Claim 1, wherein said step of filtering comprises:

c1) identifying the standard words and phrases to be contained within said thesaurus from said reference word list;

c2) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with a standard word within said thesaurus; and

c3) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with a standard phrase within said thesaurus.

9. A computer system comprising:

a bus;

a memory unit coupled to said bus; and

a processor coupled to said bus, said processor for executing a method for normalizing text in a document, said method comprising the steps of:

a) generating a list of reference words and phrases and a list of non-reference words and phrases from a selected group of documents;

b) comparing said list of reference words and phrases with a joined list containing said reference words and phrases and said non-reference words and phrases using an edit-distance algorithm to create an approximate duplicates list;

c) filtering said approximate duplicates list to create a thesaurus of standard words and phrases and their variations; and

d) editing said selected group of documents with an editor operable to use said thesaurus to replace a word or phrase on said approximate duplicates list with said standard words and phrases.

10. The computer system of Claim 9, wherein words and phrases from said selected group of documents that are on a stop word list are discarded.

11. The method of Claim 10, wherein words and phrases not discarded comprise said lists of reference and non-reference words and phrases.

12. The computer system of Claim 9, wherein said step of generating further comprises:

a1) counting the frequency of occurrence of a plurality of words and phrases from said selected group of documents;

a2) placing words and phrases with special characters embedded within them on said reference word list;

a3) processing words and phrases from said selected group of documents not already on said reference word list with a spell-checker program, wherein words and phrases that are recognized as correctly spelled are placed on said reference word list and all unrecognized words and phrases are placed on said non-reference word list;

a4) setting a frequency of occurrence threshold for said reference word list, wherein words and phrases which have a frequency of occurrence below said threshold are discarded as irrelevant; and

a5) setting a word frequency threshold for said non-reference word list, wherein words and phrases which have a frequency of occurrence above said threshold remain on said non-reference word list.

13. The computer system of Claim 12, wherein said reference word list can be merged with an existing domain specific dictionary.

14. The computer system of Claim 9, wherein said step of comparing comprises:

b1) setting parameters for said edit distance algorithm;

b2) combining said reference word list with said non-reference word list to create a joined list;

b3) comparing words and phrases on said joined list with words and phrases on said reference word list using said edit distance algorithm; and

b4) pairing words and phrases from said non-reference word list with words and phrases from said reference word list, wherein pairs of said words and phrases which are within said parameters of said edit distance algorithm are placed on said approximate duplicates list.

15. The computer system of Claim 14, wherein said step of comparing further comprises:

setting a parameter, based upon frequency of occurrence, for words and phrases not on said approximate duplicates list;

placing words and phrases not on said approximate duplicates list which are within said parameter on said reference word list; and

discarding words and phrases not on said approximate duplicates list which are outside said parameter.

16. The computer system of Claim 9, wherein said step of filtering comprises:

c1) identifying the standard words and phrases to be contained within said thesaurus from said reference word list;

c2) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with a standard word within said thesaurus; and

c3) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with a standard phrase within said thesaurus.

17. A computer-usable medium having computer-readable program code embodied therein for causing a computer system to perform the steps of:

a) generating a list of reference words and phrases and a list of non-reference words and phrases from a selected group of documents;

b) comparing said list of reference words and phrases with a joined list containing said reference words and phrases and said non-reference words and phrases using an edit-distance algorithm to create an approximate duplicates list;

c) filtering said approximate duplicates list to create a thesaurus of standard words and phrases and their variations; and

d) editing said selected group of documents with an editor operable to use said thesaurus to replace a word or phrase on said approximate duplicates list with said standard words and phrases.

18. The computer-usable medium of Claim 17, wherein words and phrases from said selected group of documents that are on a stop word list are discarded.

19. The method of Claim 18, wherein words and phrases not discarded comprise said lists of reference and non-reference words and phrases.

20. The computer-usable medium of Claim 17, wherein said step of generating further comprises:

a1) counting the frequency of occurrence of a plurality of words and phrases from said selected group of documents;

a2) placing words and phrases with special characters embedded within them on said reference word list;

a3) processing words and phrases from said selected group of documents not already on said reference word list with a spell-checker program, wherein words and phrases that are recognized as correctly spelled are placed on said reference word list and all unrecognized words and phrases are placed on said non-reference word list;

a4) setting a frequency of occurrence threshold for said reference word list, wherein words and phrases which have a frequency of occurrence below said threshold are discarded as irrelevant; and

a5) setting a word frequency threshold for said non-reference word list, wherein words and phrases which have a frequency of occurrence above said threshold remain on said non-reference word list.

21. The method of Claim 20, wherein said reference word list can be merged with an existing domain specific dictionary.

22. The computer-usable medium of Claim 17, wherein said step of comparing comprises:

b1) setting parameters for said edit distance algorithm;

b2) combining said reference word list with said non-reference word list to create a joined list;

b3) comparing words and phrases on said joined list with words and phrases on said reference word list using said edit distance algorithm; and

b4) pairing words and phrases from said non-reference word list with words and phrases from said reference word list, wherein pairs of

said words and phrases which are within said parameters of said edit distance algorithm are placed on said approximate duplicates list.

23. The computer-usable medium of Claim 22, wherein said step of comparing further comprises:

setting a parameter, based upon frequency of occurrence, for words and phrases not on said approximate duplicates list;

placing words and phrases not on said approximate duplicates list which are within said parameter on said reference word list; and

discarding words and phrases not on said approximate duplicates list which are outside said parameter.

24. The computer-usable medium of Claim 17, wherein said step of filtering comprises:

c1) identifying the standard words and phrases to be contained within said thesaurus from said reference word list;

c2) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with standard words and phrases within said thesaurus; and

c3) manually filtering said list of approximate duplicates, wherein approximate duplicates are paired with a standard phrase within said thesaurus.